37589S1

DIRECTORY SERVICE FOR FORM PROCESSING

FIELD OF THE INVENTION

The present invention relates generally to computerized information processing, and specifically to
5   extracting data from filled-in form documents.

BACKGROUND OF THE INVENTION

Methods for extraction of information filled into form documents are well known in the art. Typically, a document is printed with a form template. The template
10 contains predefined fields that are filled in by a user with appropriate characters. The document is scanned into a computer, which typically uses an optical character recognition (OCR) program to identify and code the characters in each field.

15 OCR identification of handwritten, or even typed, characters can be uncertain, due to a range of problems including uneven scan quality, variable character shapes, and interference between the filled-in characters and features of the printed template. A variety of methods
20 and systems have been developed to deal with these problems. For example, U.S. Patents 5,182,656, 5,191,525 and 5,793,887, whose disclosures are incorporated herein by reference, describe methods for registering a document image with a form template so as to remove the template
25 and extract the filled-in information from the form. Once the form is accurately registered with the known template, it is a simple matter for the computer to assign the fill-in characters to the appropriate fields. Dropping the template from the document image also
30 reduces substantially the volume of memory required to transmit or store the image.

IL9-2000-0009   1

Because of the uncertainty of machine identification of characters by OCR, methods have been developed for selectively verifying the correctness of coded results. For example, U.S. Patent 5,455,875, whose disclosure is

5 incorporated herein by reference, describes a system and method for correction of optical character recognition, based on an interactive display of OCR results that is designed to enable an operator to correct erroneous character data reliably and efficiently.

10 Even in data that are not generated by OCR, there are commonly errors and inconsistencies, such as address information that is out of date or misspelled. To deal with problems of this sort, a number of companies offer address verification services, in which a mailing list is

15 checked against an up-to-date master list. One example of such a service is "InfoBase BestAddress," offered by Acxiom Corporation, as described at www.acxiom.com. This service both identifies incorrect addresses and, where possible, provides corrections. The U.S. Postal Service

20 offers master address databases that can be used to do this sort of verification.

SUMMARY OF THE INVENTION

In preferred embodiments of the present invention, a directory service receives information extracted from a form that has been filled in by a user. The information
5   is typically sent to the directory service via a computer network by a client, who has received the filled-in form from the user and needs the information contained in one or more fields on the form to be coded and checked. The service returns coded, verified results to the client
10 over the network. Typically, multiple fields on multiple copies of the form, filled in by different users, are processed in this manner.

To deal with the information that is to be sent by the client, the directory service defines and assembles a
15 directory of data that is specific to a domain or category to which the information belongs. Preferably, the service assembles the specific directory by culling the data from other, more general databases. The service codes the information filled into the form. It then
20 looks up the coded information in the directory to check whether the information is coded correctly, to correct errors when they are detected, and/or to choose among a number of possible codes when the coding is uncertain. The use of the specific, focused directory enables the
25 service to search and check the coded information with greater reliability and speed than are generally achievable with general-purpose databases, such as public-domain telephone and address listings.

In some preferred embodiments of the present
30 invention, the users fill in the forms by writing or typing characters into the fields. Preferably, the client sends images of the filled-in field to the service

via the network, and the service uses OCR techniques to code the characters. Alternatively, the client may itself code the characters in the field and then send the coded results, or a number of alternative codes, to the

5   service. In either case, by checking the OCR output against the directory, the service is able to identify and eliminate errors in the OCR coding and to reduce the number of uncertain OCR readings that need to be passed to a human operator for verification. Thus, by using the

10   directory service, a client who is not expert in OCR and does not have convenient access to appropriate, focused directories is able to obtain high-quality coding results without a major investment in acquiring new infrastructure or capabilities.

15   Preferably, the client pays the service for providing the coded information on the basis of the quantity of information that is processed. Most preferably, the payment is calculated based upon a price per field processed. Alternatively, the payment may be

20   on the basis of processing resources, such as CPU time, expended in coding and verifying the information, or on a fixed price or subscription basis, or on substantially any other commercial basis that is known in the art.

There is therefore provided, in accordance with a

25   preferred embodiment of the present invention, a method for processing a document including a field containing information in a predefined domain, the method including:

defining a directory of data relating to the predefined domain;

30   receiving from a client via a computer network an image of the field containing the information;

processing the image to code the information; and

looking up the coded information in the directory so as to check whether the information is coded correctly.

Preferably, the method includes returning the checked, coded information over the network to the client. Most preferably, receiving the image of the field includes receiving a number of fields filled in with respective information, regarding which the checked, coded information is returned to the client, and the method includes receiving payment from the client according to the number of the fields.

Preferably, defining the directory includes selecting data specific to the predefined domain from one or more general databases.

In a preferred embodiment, receiving the image includes receiving an image of alphanumeric characters in the field. Preferably, the document includes a template delineating the field, and wherein receiving the image of the characters includes receiving the image of the characters filled into the field and remaining after drop-out of the template from the image of the field. Further preferably, processing the image includes applying computerized optical character recognition (OCR) to code the characters. Most preferably, looking up the coded information includes selecting a preferred reading of the characters from among two or more possible readings generated by the OCR, responsive to the data in the directory. Additionally or alternatively, looking up the coded information includes generating a confidence score, and processing the image includes passing the image to a human operator for coding when the confidence score is below a predetermined threshold.

Preferably, looking up the coded information includes detecting an error in the coded information and correcting the error using the data in the directory.

There is also provided, in accordance with a
5 preferred embodiment of the present invention, a method for processing forms, each form including a field that is filled in with information in a predefined domain, the method including:

defining a directory of data relating to the
10 predefined domain by selecting data specific to the domain from one or more general databases;

receiving from a client via a computer network the information that is filled into the field on the forms by a plurality of users in communication with the client;
15 and

checking whether the information is correct by looking up the information in the directory.

Preferably, receiving the information includes receiving coded information, and checking whether the
20 information is correct includes checking whether the coded information is correct.

There is additionally provided, in accordance with a preferred embodiment of the present invention, apparatus for processing a document including a field containing
25 information in a predefined domain, the apparatus including:

a memory, in which a directory of data relating to the predefined domain is stored; and

a directory service processor, adapted to receive
30 from a client via a computer network an image of the field containing the information, to process the image to code the information, and to look up the coded

37589S1

information in the directory so as to check whether the information is coded correctly.

There is additionally provided, in accordance with a preferred embodiment of the present invention, apparatus for processing forms, each form including a field that is filled in with information in a predefined domain, the apparatus including:

a memory, in which a directory of data relating to the predefined domain is stored by selecting data specific to the domain from one or more general databases; and

a processor, adapted to receive from a client via a computer network the information that is filled into the field on the forms by a plurality of users in communication with the client, and to check whether the information is correct by looking up the information in the directory.

There is moreover provided, in accordance with a preferred embodiment of the present invention, a computer software product for processing a document including a field that contains information in a predefined domain, the product including a computer-readable medium in which program instructions are stored, which instructions, when read by a computer, cause the computer to receive a definition of a directory of data relating to the predefined domain and, upon receiving from a client via a computer network an image of the field containing the information, to process the image so as to code the information and to look up the coded information in the directory so as to verify that the information is coded correctly.

IL9-2000-0009   7

There is furthermore provided, in accordance with a preferred embodiment of the present invention, a computer software product for processing forms, each form including a field that is filled in with information in a predefined domain, the product including a computer-readable medium in which program instructions are stored, which instructions, when read by a computer, cause the computer to receive a definition of a directory of data relating to the predefined domain generated by selecting data specific to the domain from one or more general databases, and upon receiving from a client via a computer network the information that is filled into the field on the forms by a plurality of users in communication with the client, to verify correctness of the information by looking up the information in the directory.

The present invention will be more fully understood from the following detailed description of the preferred embodiments thereof, taken together with the drawings in which:

37589S1

BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 is a block diagram that schematically illustrates a system for processing information filled into forms, in accordance with a preferred embodiment of

5   the present invention;

Fig. 2 is a flow chart that schematically illustrates a method for building a directory, in accordance with a preferred embodiment of the present invention; and

10   Fig. 3 is a flow chart that schematically illustrates a method for processing information filled into a form, in accordance with a preferred embodiment of the present invention.

DETAILED DESCRIPTION OF PREFERRED EMBODIMENTS

Fig. 1 is a block diagram that schematically illustrates a system 20 for processing information filled into a form 24, in accordance with a preferred embodiment

5 of the present invention. In the scenario shown in Fig. 1, a client 22, such as a system integrator, is responsible for automating data collection from a large number of forms, but does not have in house the capabilities needed to process the data automatically.

10 Rather than purchasing software and developing the necessary capabilities, which would require a large investment of time and capital, client 22 contracts with a directory service 30 to perform the processing. The directory service typically comprises one or more

15 suitable computer processors with software for carrying out the methods described hereinbelow. The software may be furnished to the directory service in electronic form, via a network or other link, or it may be supplied on tangible media, such as CD-ROM or non-volatile memory.

20 Each filled-in form received by client 22 is scanned by a scanner 26 to form an electronic image of the form, as is known in the art. The client sends the entire form image or selected elements of the image, as described hereinbelow, to service 30 via a computer network 28,

25 typically via the Internet. The directory service applies OCR to code the characters filled into the form, and then uses one or more directories 32 stored in a memory or other storage device 33 to detect coding errors and, where possible, to fix them. For example, assuming

30 form 24 to be a medical insurance form, which includes fields for the name and address of a treating physician, the directory service would preferably procure or produce

a directory of physicians against which to verify this information. After completing the coding and verification process, service 30 returns the coded results via network 28 to client 22.

5 Fig. 2 is a flow chart that schematically illustrates a method by which directory service 30 assembles the directory needed for a particular verification job, in accordance with a preferred embodiment of the present invention. Together with 10 client 22, the directory service defines a domain over which the information in form 24 is to be searched, at a search definition step 34. This domain might be the population of practicing physicians in the United States, for example.

15 At the same time, the directory service receives a definition of the specific fields that are to be coded, at a field definition step 36. In the case of the insurance form mentioned above, for example, these fields might include the physician's name, address and 20 specialization, as well as an identification of the patient and the procedure carried out. The client and directory service preferably also agree at this stage as to the form in which the field contents for processing are to be sent from the client to the service. 25 Preferably, the client sends electronic images of the fields, which are to be coded by the service using OCR. Alternatively, the field contents may be sent to the service already in coded form. This will be the case, for example, when the client itself performs the OCR 30 (thereby reducing the volume of data that must be sent over network 28) or when the forms have been filled in electronically, so that OCR is not required. Although in

this latter case the directory service no longer needs to deal with OCR coding errors, directory lookup is still useful in detecting and correcting typographical errors and other inaccuracies.

5   Based on the domain and field definitions, the directory service preferably assembles a special-purpose directory for use in verifying the results of coding the filled-in forms, at a directory building step 38. Preferably, the directory service purchases and maintains

10   a stock of specialized databases, such as the physician directory mentioned above. Alternatively or additionally, the directory service builds and maintains directories of its own, typically by assembling information from general, public-domain databases and

15   from other available sources. Further alternatively or additionally, general databases, such as postal or telephone directories, may be used when appropriate. Most preferably, the directory service employs agents and surveys sources of information to keep its directories up

20   to date.

       Fig. 3 is a flow chart that schematically illustrates a method for processing the information in form 24 by directory service 30, in accordance with a preferred embodiment of the present invention. This

25   method uses the field definitions and directory generated at steps 36 and 38, as described above. The description of the method of Fig. 3 assumes that client 22 receives paper forms, comprising a template filled in by users with handwritten or printed characters. The method is

30   also applicable, however, *mutatis mutandis,* to forms that are filled in electronically.

Each form 24 that is received by client 22 is scanned to generate an electronic image of the form, at a form input step 40. Preferably, a template registration and drop-out program, as is known in the art, is provided

5   on the client's computer in order to register the image with a template of the form and to remove the template from the image. Suitable methods for this purpose are described, for example, in the above-mentioned U.S. Patents 5,182,656, 5,191,525 and 5,793,887. Removal of

10 the template reduces the volume of information that must be transmitted over network 28 to directory service 30 and makes subsequent OCR processing easier and more accurate. Alternatively, client 22 transmits the entire image to service 30, and template drop-out is performed

15 by the service or not at all.

Following template drop-out, the fields to be coded by the directory service are located on the form, at a field identification step 44. The identification is typically based on predefined positions of the fields in

20 the form template. Preferably, this step, as well, is performed by suitable software operated by client 22, whereby only the images of the specific fields of interest are transmitted subsequently to service 30. Alternatively, the appropriate fields for processing are

25 extracted from the overall image by the directory service.

The images of the selected fields are read and coded, at a content reading step 46. Any suitable method of OCR that is known in the art may be used at this step

30 (assuming that form 24 is a paper form, whose content must be coded). Preferably, the OCR program returns one or more possible readings of the content, each with a

respective confidence score.   The results of the coding
are checked against the data in the selected directory,
at a lookup step 48.   When step 46 returned only a single
reading, step 48 is used to confirm that the coded
5   contents agree with one of the entries in the directory
(for example, · that the physician's name, address and
specialty all match).   Preferably, a "fuzzy,"
error-tolerant search algorithm is used, so that small
errors, such as misspellings or OCR misreadings, can be
10   detected and corrected, without leading to rejection of
an otherwise valid coding result.   An exemplary search
algorithm of this type is described by Wu et al., in an
article entitled, "AGREP – A Fast Approximate
Pattern-Matching Tool," published in *Proceedings of the*
15   *Winter 1992 USENIX Conference*, pages 153-162, which is
incorporated herein by reference.   When multiple,
alternate readings are suggested by step 46, the
directory lookup at step 48 is used to choose the most
likely reading among the alternatives.

20   Step 48 thus either confirms or modifies the coding
result generated at step 46.   Preferably, the confidence
score from step 46 is also modified by step 48, typically
increasing the confidence level to "certain" when an OCR
reading is found to correspond with high likelihood to an
25   entry in the directory.   On the other hand, when the OCR
reading does not correspond to any directory entry, its
confidence level may be reduced.   At a confidence
checking step 50, the confidence level of the coding
result is compared to a predetermined threshold.   If the
30   confidence is below threshold, the original field is
passed to a human operator, preferably together with the
(uncertain) coding results, at a manual coding step 52.

Any suitable method of data presentation may be used to assist the operator in processing the information efficiently and reliably, such as that described in U.S. Patent 5,455,875. The operator either confirms or

5   selects the appropriate coding result from among the alternatives offered by the OCR, or enters a different, correct result.

The verified coding result for each field is returned to client 22 at a concluding step 54.

10   Preferably, the directory service charges the client for its work on the basis of the number of fields, words or characters that have been processed. Alternatively, the charge may be based on a fixed, periodic payment, or on a measure of use of the resources of the directory service,

15   such as CPU time, or on substantially any other payment basis known in the art.

While preferred embodiments described herein relate particularly to form documents and OCR coding, it will be understood that the principles of the present invention

20   are similarly applicable to error checking, correction and verification of data coding generated by other methods and to processing documents of other types. It will thus be appreciated that the preferred embodiments described above are cited by way of example, and that the

25   present invention is not limited to what has been particularly shown and described hereinabove. Rather, the scope of the present invention includes both combinations and subcombinations of the various features described hereinabove, as well as variations and

30   modifications thereof which would occur to persons skilled in the art upon reading the foregoing description and which are not disclosed in the prior art.